KOLMOGOROV-SMIRNOV TEST FOR GROUPED DATA

Suharto Darmosiswoyo

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

KOLMOGOROV-SMIRNOV TEST FOR GROUPED DATA

by

Suharto Darmosiswoyo

March 1975

Thesis Advisor: Donald R. Barr

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

Kolmogorov-Smirnov Test for Grouped Data

5. TYPE OF REPORT & PERIOD COVERED
Master's Thesis
March 1975

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Suharto Darmosiswoyo

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Naval Postgraduate School
Monterey, California 93940

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Naval Postgraduate School
Monterey, California 93940

12. REPORT DATE
March 1975

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Thesis Advisor: Associate Professor D. R. Barr
Autovon: 479-2654

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Grouped data
Conover's procedure
K-S test for grouped data
Class boundary

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Herein is presented a Kolmogorov-Smirnov test for grouped samples. It constitutes an application of W.J. Conover's procedure, which was originally designed for calculating exact critical levels for test with discrete distributions. This test extends the goodness-of-fit test to samples which are grouped into intervals. Critical levels in the two-sided case were calculated to a close approximation.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |
(Page 1)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Some examples of the application of this extension of the Kolmogorov-Smirnov test are included and comparisons with the Chi-square test, the Kolmogorov-Smirnov test for continuous distribution and for grouped samples are made. Comparisons among the three goodness-of-fit tests based on the results of computer simulations are made.

Kolmogorov-Smirnov Test For Grouped Data

by

Suharto Darmosiswoyo

Lieutenant Colonel, Indonesian Army

Submitted in partial fulfillment of the
requirement for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1975

ABSTRACT

Herein is presented a Kolmogorov-Smirnov test for grouped samples. It constitutes an application of W.J. Conover's procedure, which was originally designed for calculating exact critical levels for test with discrete distributions. This test extends the goodness-of-fit test to samples which are grouped into intervals. Critical levels in the two-sided case were calculated to a close approximation.

Some examples of the application of this extension of the Kolmogorov-Smirnov test are included and comparisons with the Chi-square test, the Kolmogorov-Smirnov test for continuous distribution and for grouped samples are made. Comparisons among the three goodness-of-fit tests based on the results of computer simulations are made.

TABLE OF CONTENTS

I. INTRODUCTION

One of the most important tests in statistical applications arises in testing hypotheses about the distribution of a population. Before specifying a model, usually we should look at the data to see if they appear to have come from the distribution which we expect to use for the model. This can be approached through the histogram, which gives us information about the density function of the underlying distribution. Another approach is the sample distribution function, which gives us an estimate of the underlying cumulative distribution function.

We call a test "a test of goodness-of-fit" if the test is concerned with the agreement between the distribution of a set of sample values and some theoretical distribution.

Much work has been devoted to finding test statistics whose distributions do not depend on parameters in the distribution of the underlying population. Such tests are commonly called distribution-free tests.

One of the most well-known and useful goodness-of-fit tests is the Kolmogorov-Smirnov Goodness-of-fit tests . The Kolmogorov-Smirnov test treats individual observations separately and thus does not lose information through grouping as does the Chi-square test. Consequently, in the continuous case, the Chi-square test is frequently less powerful than the Kolmogorov-Smirnov test.

It is also known that the Kolmogorov-Smirnov test is conservative when the hypothesized distribution function is not continuous. In many situations, observations from a continuous distribution are grouped. However, studies of the modifications of the Kolmogorov-Smirnov test for use with grouped data appear to be rather limited.

The Chi-square test, suggested by Pearson (1900), is well suited for use with grouped data, whereas the K-S test is for random samples from continuous populations. W.J. Conover's procedure was designed to calculate critical levels for random samples from discrete distributions. Since grouping the data from continuous or discrete populations will result in a corresponding underlying distribution which is discrete, Conover's procedure can also be used for grouped data. The general problem of defining the classes or determining the class boundaries in some optimal way has apparently received limited attention. In what follows we propose a Kolmogorov-Smirnov test for grouped data which shows a method of finding (almost) exact critical levels and the power of this test. Thus the Kolmogorov-Smirnov test may be used as a goodness-of-fit test, regardless of whether the hypothesized distribution is continuous, or whether the samples are grouped.

The following listing constitutes the description or definition of notation used herein:

Notation	Description
K-S test	Kolmogorov-Smirnov test for continuous distribution function and ungrouped data.
$K\text{-}S_{(g)}$ test	Kolmogorov-Smirnov test for grouped data.
S_n	Empirical distribution function of a random sample of size n
S_n^g	Empirical distribution function of a random sample of size n which is grouped.
n	Sample size.
α	Critical level of test
δ	Critical value associated with continuous distribution function and ungrouped data.
δ_g	Critical value associated with grouped data.
F_o	Hypothesized distribution function.
F_x	Some population distribution function
b_i	Some fixed real number representing the class boundary.
H_o	Null hypothesis
H_1	Alternative hypothesis
p_i	Relative frequency of acceptance of null hypothesis.
$x_{(1)} \leq x_{(2)} \leq \cdots x_{(n)}$	Ordered observed random sample from distribution F_x

II. GENERAL DESCRIPTION OF THE TEST

The Kolmogorov-Smirnov test for goodness-of-fit is based on the maximum vertical difference between the empirical and the hypothesized cumulative distribution function. Under H_o, the empirical distribution function is expected to be close to the specified distribution function, $F_o(\cdot)$. If the maximum vertical difference between the specified and the empirical distribution is not small enough for some x, say greater than a critical value δ, this may be considered evidence that the hypothesized distribution is not the one from which the sample was drawn.

Suppose $F_o(\cdot)$ is the hypothesized distribution and $S_n(\cdot)$ is the empirical distribution function, then

$S_n(x) = \frac{k}{n}$ for x between the $k^{\underline{th}}$ and (k + 1) st largest values in the sample and n is the sample size.

The Kolmogorov-Smirnov test is: accept the hypothesis H_o: $X \sim F_o(\cdot)$ if and only if

$$D_n = \sup_{-\infty<x<\infty} \left| F_o(x) - S_n(x) \right| \leqslant \delta \text{ , where}$$

δ is adjusted to give a level α test. This is sometimes called the two-sided Kolmogorov-Smirnov statistic as opposed to the one-sided statistics D_n^+ and D_n^-, where

$$D_n^+ = \operatorname{Sup}_{-\infty<x<\infty} \left\{ S_n(x) - F_o(x) \right\}$$

$$D_n^- = \operatorname{Sup}_{\infty<x<\infty} \left\{ F_o(x) - S_n(x) \right\}$$

It is well know that if $X_1, X_2, ---, X_n$ is a random sample from a continuous distubition function $F_o(\cdot)$, D_n is distribution-free i.e.; has distribution independent of $F_o(\cdot)$.

Let $X_1, X_2, ---, X_n$ be independent random variables with the common cumulative distribution function $F_o(\cdot)$. Furthermore let $x_{(1)} \quad x_{(2)} \quad --- \quad x_{(n)}$ be the result of n independent observations, arranged in order size, that is, the n order statistics. Suppose the above random samples are divided into k groups, given by class boundaries x_{j_i}, i=1, 2, --- k, where the b_j's are the outcomes (observed values) of certain of the order statistic $X_{(j)}$.

Let $S_n^g(x)$ be the sample distribution function based on the grouped data. Then $S_n^g(x)$ is a step function which jumps at class boundaries and so can be written;

$$S_n^g(x) = \begin{cases} o & \text{for } x < X_{(1)} \\ \dfrac{j_i}{n} & \text{for } X_{(j_i)} \leqslant x < X_{(j_{i+1})} \\ 1 & \text{for } x > X_{(n)} \end{cases}$$

For the grouped data, the maximum vertical difference between $F_o(x)$ and $S_n^g(x)$ occurs at one of the class boundaries. The two-sided Kolmogorov-Smirnov test statistics for grouped data ($K - S_{(g)}$ statistic) is defined to be;

$$D_n = \max_{x \in \{x_{j_1} \cdots x_{j_{k-1}}\}} \left| F_o(x) - S_n^g(x) \right|$$

The "one-sided $K - S_{(g)}$ test" statistics are;

$$D_n^+ = \max_{x \in \{x_{j_1} \cdots x_{j_{k-1}}\}} \left\{ S_n^g(x) - F_o(x) \right\}$$

$$D_n^- = \max_{x \in \{x_{j_1} \cdots x_{j_{k-1}}\}} \left\{ F_o(x) - S_n^g(x) \right\}$$

In what follows, it will be shown that these three statictics are distribution free, provided that $F_o(x)$ is continuous.

Make the change of variables;

$y = F_o(x_{j_i})$

$Y_j = F_o(X_j)$. Since $F_o(x_{j_i})$ is nondecreasing, the inequality $X_i \leq x_{j_i}$ is equivalent to $F_o(x_i) \leq F_o(x_{j_i})$ or $Y_i \leq y$. Thus $S_n^g(x_{j_i}) = \frac{j_i}{n} = \frac{1}{n}$ times (the number of $X_j \leq x_{j_i}$) $= \frac{1}{n}$ (the number of $Y_j \leq y$) $= H(y)$

Furthermore, the distribution of Y_j is;

$P\left[Y_j \leq y\right] = P\left[X_i \leq x_{j_i}\right] = F_o(x_{j_i}) = y$. Thus,

(This page intentionally blank)

Y_j has a uniform distribution and D_n can be written as

$$D_n = \max_{x \in \{x_{j_1} \text{---} x_{j_{k-1}}\}} \left| F_o(x) - S_n^g(x) \right|$$

$$= \max_{i=1,\ 2\text{---}k} \left| F_o(x_{j_i}) - S_n^g(x_{j_i}) \right|$$

so that, $\left| F_o(x_{j_i}) - S_n^g(x_{j_i}) \right| = \left| y - H(y) \right|$

and $D_n = \max_{x \in \{x_{j_1} \text{---} x_{j_{k-1}}\}} \left| F_o(x) - S_n^g(x) \right| = \max_{y \in \{y_1 \text{---} y_{k-1}\}} \left| y - H(y) \right|$

The expression on the right is the vertical distance between the sample distribution of Y_1, Y_2 ----Y_n each U(0,1). Since this expression does not depend upon F_o, D_n is distribution free. In this case, where the class boundaries are order statistic, D_n is distribution free. Massey's procedure (8) or Davidson's method (4) could be used to calculate critical values for this test.

In the case where the class boundaries are fixed constant, D_n is no longer distribution free. It can be described as follows; Suppose the class boundaries are fixed constants in advance say b_i. The value of $F_o(\cdot)$ at b_i is $F_o(b_i)$. As before, make the change of "variable" $y = F_o(b_i)$. However, since b_i is constant, $y = F_o(b_i)$ cannot be transformed to $Y = F_o(B_i)$. In other words D_n =

$$\max_{x \in \{b_i\}} \left| F_o(x) - S_n^g(x) \right| = \max_i \left| F_o(b_i) - S_n^g(b_i) \right|$$

$\neq \max \left| y - H(y) \right|$. Therefore, it is suggested that Conover's procedure be used to calculate critical levels for $K - S_{(g)}$ when the class boundaries are fixed constant.

III. TEST PROCEDURE AND CRITICAL LEVEL

A. TEST-PROCEDURE

Unlike the Chi-square test and the Kolmogorov-Smirnov test, in which critical values are calculated which correspond to selected critical levels, with Conover's procedure one calculates critical levels which correspond to selected critical values. The procedure is as follows. Let $X_1, X_2 \text{---} X_n$ be a random sample of size n, drawn from some unknown discrete population distribution $F_x(\cdot)$. The hypothesis is: $H_o: F_x(\cdot) = F_o(\cdot)$ where $F_o(\cdot)$ has all parameters specified. The alternative is $H_1: F_x(\cdot) \neq F_o(\cdot)$. The test statistic, either D_n, D_n^+, or D_n^- depending on which is desired, is calculated. Further, the critical level is computed and if the value of this critical level is greater than or equal to that specified in advance (usually either 005 or 001), the null hypothesis is accepted; otherwise it is rejected.

Let δ be some fixed real number, $o < \delta < 1$. The critical value of the test, that is $P\left[D_n \geq \delta\right]$ for the two-sided Kolmogorov-Smirnov (discrete) test, can be computed by using a procedure due to W.J. Conover (2). Infact this is an approximation method, since $P\left[D_n \geq \delta\right]$ is obtained by calculating $P\left[D_n^- \geq \delta\right]$ and $P\left[D_n^+ \geq \delta\right]$, where $\delta = \max(\delta^-, \delta^+)$ and δ^-, δ^+ are the observed values of D_n^- and D_n^+ respectively. $P\left[D_n \geq \delta\right]$ is taken to be approximately

approximately $P\left[D_n^- \geq \delta\right] + P\left[D_n^+ \geq \delta\right]$. To see the adequacy of this approximation, we proceed as follows;

$$P\left[D_n \geq \delta\right] = P\left[(D_n^+ \geq \delta) \cup (D_n^- \geq \delta)\right]$$

$$= P\left[D_n^+ \geq \delta\right] + P\left[D_n^- \geq \delta\right] - P\left[(D_n^+ \geq \delta) \cap (D_n^- \geq -\delta)\right]$$

$$P\left[(D_n^+ \geq \delta) \cap (D_n^- \geq \delta\right] = P\left[D_n^+ \geq \delta\right] P\left[D_n^- \geq \delta \mid D_n^+ \geq \delta\right] \cdots (1)$$

Since $P\left[D_n^- \geq \delta \mid D_n^+ \geq \delta\right] \leq P\left[D_n^- \geq \delta\right]$ equation (1) becomes:

$$P\left[(D_n^+ \geq \delta) \cap (D_n^- \geq \delta)\right] \leq P\left[D_n^+ \geq \delta\right] P\left[D_n^- \geq \delta\right]$$

an approximate value of $P\ D_n \geq \delta$ is thus

$P\left[D_n \geq \delta\right] \approx P\left[D_n^+ \geq \delta\right] + P\left[D_n^- \geq \delta\right]$. The error of this approximation is $P\left[(D_n^+ \geq \delta) \cap (D_n^- \geq \delta)\right]$ which is less than the product $P\left[D_n^+ \geq \delta\right] P\left[D_n^- \geq \delta\right]$

But in practice, δ is taken so that $P\left[D_n^+ \geq \delta\right]$ and $P\left[D_n^- \geq \delta\right]$ are small and then the product is quite small and can be safely ignored. For example, suppose both one-sided tests have the same critical level, say, 0.05, then $P\left[D_n \geq \delta\right] \approx .10$ with an error less than (0.05) (0.05) = 0.0025.

B. CRITICAL-LEVEL

Let $\delta^- = \max_{-\infty<x<\infty} (F_o(x) - S_n(x))$ where $S_n(x)$ represents the empirical cumulative distribution function of sample (Fig. 1)

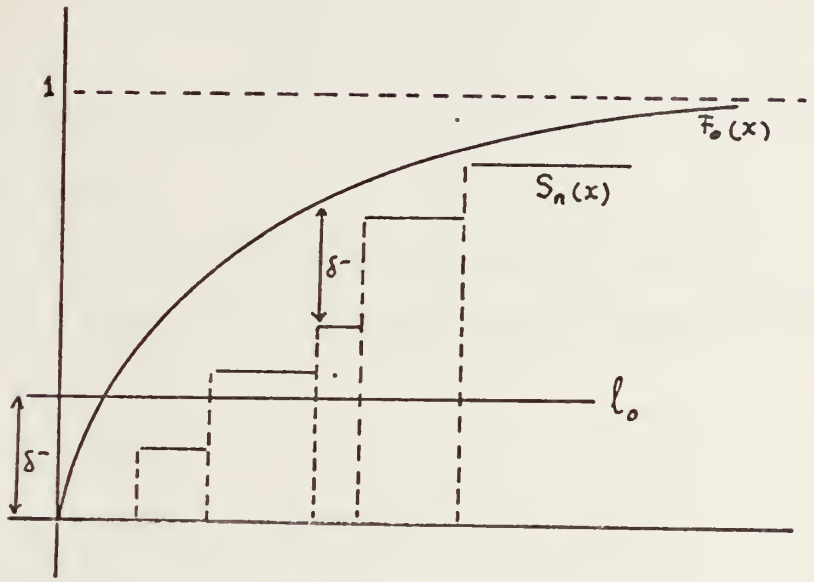

Figure 1

δ^- is the maximum vertical difference between the expected cumulative distribution function and the observed cumulative distribution function. Draw horizontal lines ℓ_j with intercept $\delta^- + \frac{j}{n}$ where $0 \leq j \leq n(1-\delta^-)$. Then compute the value of $f_j = 1 - (\delta^- + \frac{j}{n})$. For $j=0$ the line ℓ_o has intercept δ^-. The value of f_j is $f_o = 1-\delta$. For $j=1$ the intercept of ℓ_1 is $(\delta^- + \frac{1}{n})$ and $f_1 = 1-(\delta^- + \frac{1}{n})$ etc.

If the horizontal lines ℓ_j intersects $F_o(\cdot)$ at a jump point (at a point x of discontinuity) the value of f_j is 1 minus the ordinate $F_o(x)$ at the top of jump.

As shown by Conover (2), the critical level $P\left[D_n^- \geq \delta^-\right]$ is;

$$P\left[D_n^- \geq \delta^-\right] = \sum_{j=0}^{n(1-\delta^-)} \binom{n}{j} f_j^{n-j} C_j \quad \text{----------(2)}$$

where

$$C_j = 1 - \sum_{j=0}^{k-1} \binom{k}{j} f_j \quad \text{----------(3)}$$

k is defined to be the largest value of the subscript j such that $f_j > 0$.

Conover's derivation and proof are tedious and are omitted here. For example, for k=5 equation (3) becomes;

$$C_o = 1$$

$$C_1 = 1-f_o$$

$$C_2 = 1-f_o^2 - 2f_1C_1$$

$$C_3 = 1-f_o^3 - 3f_1C_1 - 3f_2C_2$$

$$C_4 = 1-f_o^4 - 4f_1^3C_1 - 6f_2^2 - 4f_3C_3$$

To calculate $P\left[D_n^+ \geq \delta^+\right]$, where $\delta^+ = \max_{-\infty<x<\infty} \left\{S_n(x) - F_o(x)\right\}$, draw horizontal lines ℓ_j with intercepts $1 - (\delta^+ + \frac{j}{n})$. The value of f_j is $1 - (\delta^+ + \frac{j}{n})$, but if the horizontal line ℓ_j intersects $F_o(\cdot)$ at a jump x, then the value of f_j is the height of $F_o(x)$ at the bottom of the jump. For this critical value, equation (2) becomes:

$$P\left[D_n^+ \geq \delta^+\right] = \sum_{j=0}^{n(1-\delta^+)} \binom{n}{j} f_j C_j \quad \text{----------(4)}$$

where c_j is given in equation (3).

Suppose we have a sample of size n grouped into k intervals, having m_i observation for the $i^{\underline{th}}$ intervals, so $\sum_{L=1}^{k} m_i = n$. Let b_i be a fixed class boundary of the $i^{\underline{th}}$ interval and $S_n^g(\cdot)$ be the empirical cumulative distribution function of the grouped sample. $S_n^g(\cdot)$ is a step function which rises by jumps of d_i where $0 \leq d_i \leq 1$,

For example, for i=1, the jump occurs at b_1 with the value of d_1. The first jump of $S_n^g(\cdot)$ occurs at the first class boundary. For i=2, the second class boundary is b_2 and at this point the value of the step function is $(d_1 + d_2)$. So, the step function $S_n^g(\cdot)$ can be written as:

$$S_n^g(x) = \begin{cases} 0 & \text{for} \quad x < X_{(1)} \\ \sum_{i=1}^{j} d_i & \text{for} \quad b_j \leq x < b_{j+1} \\ 1 & \text{for} \quad x > X_{(n)} \end{cases}$$

where $$\sum_{i=1}^{j} d_i = \frac{\text{the number of } X_{(i)} \leq b_j}{n}$$

In using Conover's procedure for K-$S_{(g)}$, let δ_g^- be the maximum vertical distance between $F_o(\cdot)$ and $S_n^g(\cdot)$, that is

$\delta_g^- = \underset{x \in \{b_1 \cdots b_k\}}{\text{Max}} (F_o(x) - S_n^g(x))$, or it can be written (Fig. 2) as;

$$\delta_g^- = \underset{i}{\text{Max}} (F_o(b_i) - S_n^g(b_i))$$

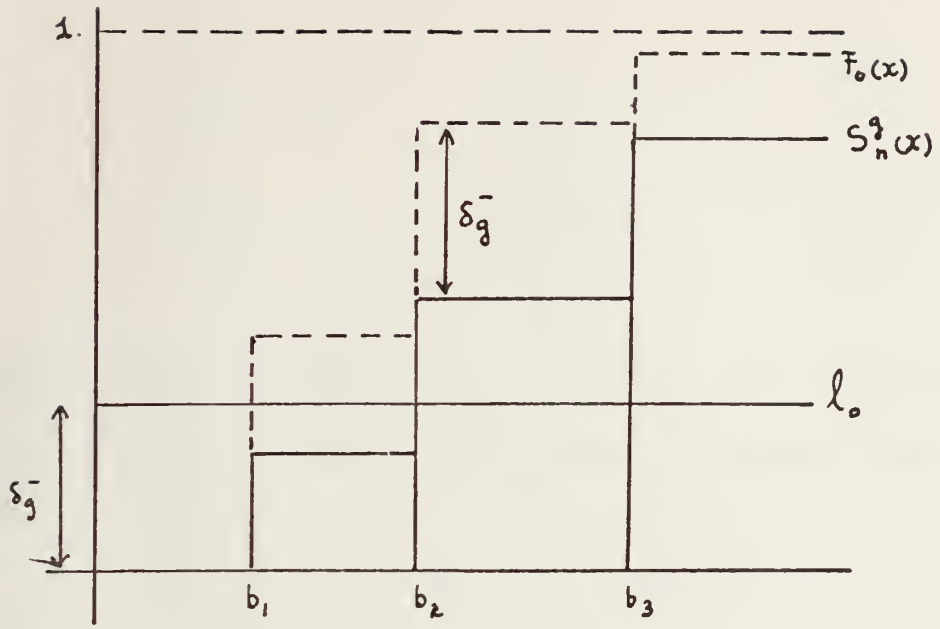

Figure 2

As before, draw horizontal line ℓ_j and then compute the values of f_j. The values of C_j are calculated by using equation (3) and finally the critical level computed by equation (4).

As the sample size is increased, the calculation (if done by hand) will be more tetious, so this procedure should be used for small sample sizes, say less than 30. Since the critical level is a decreasing function of δ_g^+ and δ_g^-, the bigger the value of this maximum vertical difference between $F_o(\cdot)$ and $S_n^g(\cdot)$ the smaller will be the critical level, and vice versa. For illustration the following is an example of how to use Conover's procedure. Suppose a random sample of size 16 is drawn from some population $F_X(\cdot)$. The observed values are (ordered):

10,35	12,04	14.16
10.40	12.53	14.26
10.42	12,63	14,40
11,46	13,18	14,47
11.49	13,33	14.84
	13.54	

The hypothesis is H_o: $X \sim U(10,20)$ with alternative H_1: $X \not\sim U(10,20)$. The samples are grouped into three intervals. The values of the class boundaries choosen to be $b_1=12$, $b_2=14$ so the three intervals are: $(-\infty\ -12)$ $(12 - 14)$ and $(14 - \infty)$. (Fig. 3)

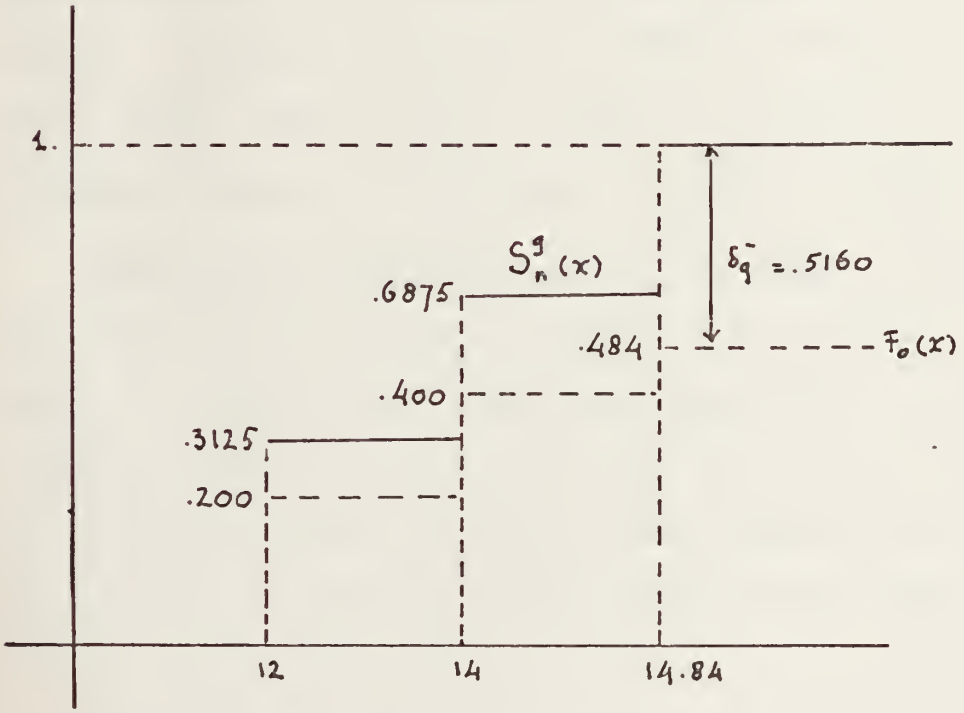

Figure 3

The values of the step function $S_n^g(\cdot)$ at the first class boundary is $S_n^g(x) = \frac{1}{16}$ (the number of $X_i \leqslant 12$) = .3125, at the second class boundary is $S_n^g(x) = \frac{1}{16}$ (the number of

$X_i \leq 14) = .6875$. Under the null hypothesis, the theoretical rates for these three intervals are $\frac{12 - 10}{10} = .20$, $\frac{14 - 10}{10} = .40$ and $\frac{14.84 - 10}{10} = .484$ respectively.

Here the maximum vertical difference is $1 - .484 = .5160$. The ℓ_0 with intercept .5160 intersects $F_0(\cdot)$ at jump point so $f_0 = 1 - .6875 = .3125$, ℓ_1 with intercept $.5160 + \frac{1}{16} = .5785$ also intersects $f_0(\cdot)$ at jump point so $f_1 = .3125$ and finally the intercept of ℓ_2 is .6410, so $f_2 = .3125$. The value of k in formula (3) is equal to 2 since ℓ_3 intersects $F_0(X)$ at a jump point with top ordinate 1.00, so $f_3=0$ which is not considered. By using formula (3) C_j is computed and found to be $C_0=1$, $C_1 = .6875$, $C_2 = .4727$. Thus, $P\left[D_n^- \geq .5160\right] = (.3125)^{16} + \binom{16}{1}(.3125)^{15}(.6875) + \binom{16}{2}(.3125)^{14}(.4727) = .000005$.

Since $F_0(x)$ is symmetric $P\left[D_n^+ \geq .5160\right]$ is equal to the above, so that $P\left[D_n \geq .5160\right] = 2 \times .000005 = .00001$. Using $\alpha = .01$ the null hypothesis should be rejected since $0.00001 < 0.01$. Incidentally this is a correct decision in this case, since the above random sample was actually drawn from a population having distribution U(10, 15).

IV. COMPARISON AND THE K-S$_{(g)}$ TEST POWER

The basic difference among the three tests, the Chi-square, the K-S and the K-S$_{(g)}$ is that the Chi-square test is sensitive to vertical deviation between the observed and expected histogram, the K-S test is based on vertical deviation between the observed and expected cumulative distribution function, whereas K-S$_{(g)}$ is based on vertical deviation between the observed and expected cumulative distribution function associated with discrete groups. Another obvious difference is that K-S statistic is distribution free, whereas for K-S$_{(g)}$, the statistic is not distribution free.

Both Chi-square and K-S$_{(g)}$ require that the data be grouped; in contrast, the K-S test does not. Therefore when the underlying distribution is continuous the K-S test permits us to investigate the goodness-of-fit with information from each observation. By contrast, both Chi-square and K-S$_{(g)}$ lose some information since individual observations are grouped into a relatively small number of classes. Further, the Chi-square and K-S$_{(g)}$ tests are affected by the number and the length of the class intervals which are choosen arbitrarily by the experimenter. The following is another example of applying the K-S$_{(g)}$ which is followed by comparison with Chi-square test. Suppose we have a

random sample of size n=15, drawn from some population $F_X(\cdot)$. They are (ordered);

2.01	3.91	14.89
2.08	5.32	15.27
2.24	9.09	21.34
2.52		29.73
2.70		36.54
		39.43
		40.74

The hypothesis to be tested is that $S_n(x)$ has come from exponential distribution with $\lambda=6$ at $\alpha=0.05$.

H_o: $X \sim$ EXP (6)

H_1: $X \not\sim$ EXP (6)

Suppose the sample had been divided into three groups associated with the intervals (0, 2.70), (2.70, 9.09) and (9.09, ∞). Under the null hypothesis the expected grouped c.d.f. has ordinates; .3624, .7791 and 1 respectively. The observed and expected frequencies for each group are 5, 3, 7 and 5.436, 6.2505, 3.3135 respectively.

The maximum vertical difference δ_g occurs at x = 9.09 where δ_g = .7791 - .5333 = .2458. The values of f_j and C_j are computed as in the previous example and are found to be;

$f_0 = .6376$	$f_7 = .2209$	$C_0 = 1$	$C_7 = .1585$
$f_1 = .6376$	$f_8 = .2209$	$C_1 = .3624$	$C_8 = .1287$
$f_2 = .2209$		$C_2 = .1314$	
$f_3 = .2209$		$C_3 = .2117$	
$f_4 = .2209$		$C_4 = .2335$	
$f_5 = .2209$		$C_5 = .2197$	
$f_6 = .2209$		$C_6 = .1914$	

With those values of f_j and C_j, $P\left[D_n^- \geq .2458\right]$ is found to be .0395. Using a similar procedure it can be seen that $P\left[D_n^+ \geq .2458\right] = .0.0162$, so $P\left[D_n \geq .2458\right]$ is approximately 0.0395 + 0.0162 = 0.0557. The hypothesis is accepted at $\alpha = 0.05$.

The critical value of the Chi-square statistic on the same sample, with three groups and hence 2 degrees of freedom, is 5.3936. This is less than 5.99, the χ^2 critical value for $\alpha = 0.05$. Again the hypothesis would be accepted, that is both tests accept that the random sample has been obtained from a population having exponential distribution with $\lambda=6$. Using interpolation in tables of the incomplete gamma function, the critical value of the Chi-square statistic associated with $\alpha=0.0557$ is found to be approximately 5.7724. For $\alpha=0.06$ this approximate value is 5.6275. If we had used $\alpha=0.06$ rather than 0.05 the $K\text{-}S_{(g)}$ test would have rejected the null hypothesis since $0.0557 < 0.06$. However the Chi-square test would have accepted the null hypothesis since 5.3936 is less

than 5.6275. The Chi-square test will reject the null hypothesis at any α level which is greater than 0.06.

In fact, the random sample for the present examples was generated from a population having an exponential distribution with mean 15. Although the critical level $\alpha=0.06$ is not commonly used, the Chi-square cannot detect that the random sample did not come from exponential distribution with mean 6.

From the example in Chapter III, it was found that the null hypothesis was rejected at $\alpha=0.01$. With the same data, with three groups, hence 2 degrees of freedom, under the null hypothesis the value of the Chi-square test statistic is 13.4075. This is greater than 10.6, the $\chi^2_{(2)}$ value at $\alpha=0.05$. Thus, both tests reject the null hypothesis at $\alpha=0.05$. The critical value of the Chi-square test associated with $\alpha=0.00001$ is found to be approximately 23.4, which means that if we had used $\alpha=0.00001$, the Chi-square test would have accepted the null hypothesis, since $13.4075 < 23.4$. Of course, the critical level $\alpha=0.00001$ is rarely used in practice. The acceptance to the null hypothesis by the Chi-square test at $\alpha=0.00001$ is most likely caused by the fact of very low value of the expected frequencies in the fourth group of the sample that is 1.344.

Suppose the null hypothesis is H_o: $X \sim U(10, 18)$ with critical level $\alpha=0.05$. Under this null hypothesis, the value of the Chi-square statistic is 5.320, which is less than 10.6. In this case, the Chi-square test accepts the

null hypothesis, or in other words, the Chi-square test accept that the random sample was generated from a population having U(10, 18) distribution. Again in this case the Chi-square test has low expected frequencies in the fourth group namely 1,92.

For the $K\text{-}S_{(g)}$ test, under the null hypothesis U(10, 18), the value of δ_g is .395, so $P\left[D_n \geq .395\right] =$ 0.00008, which is less than 0.05. The $K\text{-}S_{(g)}$ test rejects the null hypothesis, which means rejects that the random sample was generated from a population having U(10, 18) distribution.

The present examples suggest the $K\text{-}S_{(g)}$ may be more powerful than the Chi-square test, at least in certain cases. Simulations conducted to explore this question are discussed below.

In the previous example, concerning the exponential distribution, the critical value of the K-S statistic in the continuous case at α=0.0557 is between .304 and .338, which leads us to reject the null hypothesis that the sample was generated from an EXP (λ=6) distribution. At the same critical level, for example 0.0557 or 0.00008, grouping samples into intervals tends to lower the power. Thus, for grouped data (samples), the appropriate critical values are smaller than those tabulated for continuous case. Thus, use of tables of critical values of the K-S test will not give a correct test. Therefore the need for the $K\text{-}S_{(g)}$ procedure is evident.

V. SIMULATION

In order to get a better comparison between the three goodness-of-fit tests previously mentioned, especially their powers, simulation was used. The powers of the tests should be compared under the same conditions, namely at the same significance level and for the same null hypothesis. There are two procedures for this simulation. The first procedure is as follows: generate a random sample of size n from a population having distribution $F_x(\cdot)$. Specify a critical level α, for example α_1 and find the critical value δ_1 of the K-S statistic associated with α_1 and n. By using tables of the incomplete gamma function, one can calculate (at least approximately) the critical values of the Chi-square statistics associated with α_1. Let C_1 denote this critical value. These three values, α_1, δ_1, C_1, are then used as the input to the simulation for computing the relative frequency of acceptance of H_o by the $K\text{-}S_{(g)}$, the K-S for continuous distribution and the Chi-square tests, respectively. Further, generate the random samples N times. The hypothesis to be tested is: H_o: $F_x(\cdot) = F_o(\cdot)$ and H_1: $F_x(\cdot) \neq F_o(\cdot)$ at level α_1.

Let $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$ denote the relative frequency of acceptance of H_o by the K-S, the $K\text{-}S_{(g)}$ and the Chi-square respectively, then

$$\hat{p}_1 = \frac{\text{"number of times"}\ (D_n \leq \delta_1)}{N}$$

$$\hat{p}_2 = \frac{\text{"number of times"}\ \ P\left[D_n \geq \delta_1\right] \geq \alpha_1}{N}$$

$$\hat{p}_3 = \frac{\text{"number of times"}\ (\chi^2_{(r)}\ \text{Statistic} \leq C\ )}{N}$$

Both acceptances of K-S and Chi-square test can be programmed. However, it appears to be difficult to program Conover's procedure. To the knowledge of the author, a subroutine for this procedure is not available.

Therefore the simulation has been done indirectly, using a second procedure, This can be described as follows: With the same random samples from the same $F_o(\cdot)$ and the same number of groups then for $0 < d_1 \leq d_2 < 1$, $P\left[D_n \geq d_1\right] \geq P\left[D_n \geq d_2\right]$. If the hypothesis is accepted at $P\left[D_n \geq d_2\right] = \alpha_2$, then it will be accepted at $P\left[D_n \geq d_1\right] = \alpha_1$, since $\alpha_1 \geq \alpha_2$. In other words the sample rates of acceptance of $K\text{-}S_{(g)}$ is defined to be

$\frac{\text{"number of times"}\ (D_n \leq d)}{N}$ where d is specified in advance. Once d is fixed, $P\left[D_n \geq d\right] = \alpha_1$ can be computed from which the critical values for the other two tests associated with α_1 can be calculated. However, it will be difficult to obtain an approximate value for the critical values of the K-S test because of the limited α values in the tables. Therefore, d is used to calculate the frequency

of acceptance by both K-S tests. So, the inputs for the second procedure are C_1 and d.

For the purpose of this simulation, groups of 15 random numbers were generated from Exponential distributions with means λ=6, 9, 12 and 15; random samples of size 14 were generated from Normal distributions with μ=0, σ=1, 3 and 5; random samples of size 16 were generated from Uniform distributions U(10, 13), U(10, 15) and U(10, 17). The value of d was choosen arbitraryly to be d=.3. The results of this simulation are summarized in table (5.1).

The null hypothesis for the four Exponential cases was F_o = EXP (λ=9), with 4 groups given by the intervals (0 -3), (3 - 6), (6 - 9) and (9 - 100). By Conover's procedure $P\left[D_n \geq d=.3\right] = \alpha_1^e = 0.0328$. The null hypothesis for the three Normal cases was N(0, 1). Four groups were used with intervals $(-\infty, -1)$, (-1, 0), (0, 1) and $(1, \infty)$ and $P\left[D_n \geq d=.3\right] = \alpha_1^n$ was found to be 0.0185. Finally, the null hypothesis for the Uniform cases was U910, 15), with 5 groups given by the intervals (10, 11), (11,. 12), (12, 13), (13, 14) , (14, 17) and $P\left[D_n > d=.3\right] = \alpha_1^u$ was found to be 0.022. The associated critical values for the Chi-square test are approximately 8.7517, 8.00, 13.245 in the EXPONENTIAL, NORMAL and UNIFORM cases, respectively. These three critical values together with the three α levels are then used in the input to the simulation.

Simulation results give p_2=97.97% for the acceptance frequency of the $K\text{-}S_{(g)}$ test for the null F_o = U(10, 15)

Random Samples	Size	d	Number groups	$P[D \geq d] = \alpha$	Critical region of $X^2_{(n)}$ associated with α	Null hypothesis	% of acceptance of K-S Test		
							$X^2_{(n)}$	Group	Continuous
N (0,1)	14	.30	4	.0185	10.50*	N (0,1)	98.7400	98.2300	87.1700
N (0,9)	14						27.2500	54.7900	16.8700
N (0,25)	14						6.0700	28.7300	2.4500
EXP (6)	15						87.2200	77.4500	65.7300
EXP (9)	15	.30	4	.0328	8.7517	EXP (9)	96.0400	96.4400	88.9650
EXP (12)	15						94.9000	93.9100	74.1600
EXP (15)	15						87.6800	84.6800	51.9300
U (10,13)	16						47.04	0	0
U (10,15)	16	.30	5	.022	13.245	U (10,15)	98.8500	97.9700	90.8200
U (10,17)	16						77,9200	75.7400	34.3200
U (0,18)	15						90.8100	88.9800	67.7000
U (0,20)	15	.30	4	.0328	8.7517	EXP (9)	86.1100	82.1000	54.4600
U (0,22)	15						97,0900	73.500	42.7700

* has been adjusted from 8.00 to 10.50

Table 5.1

distribution, which mean (1.00 - .9797) = 0.0203 is the estimated rejection rate, in other words α_1=0.0203. The input α_1 for the Uniform random cases is 0.022; the difference 0.0017 between these is not significant. From table (5.1) it can be seen that the corresponding difference for the Exponential case is 0.0028 and for the Normal case is 0.0008, which are not significant. From the same table it can be seen that the significance level (for the null hypothesis) for both the Chi-square and K-$S_{(g)}$ tests in the Uniform and Exponential cases, are very close.

For example, in the Exponential case, this difference is only (.0356 - .0316) = .004. Therefore these tests can be assumed to be at the same α level. In the Normal case by grouping the data, the value of α_1=0.0177. This gives a difference of 0.0008 compared with the input α_1=0.0185. However, the critical value of the Chi-square statistic associated with α_1=0.0185, and degrees of freedom 3 is 8.0 which gives .9565 for the acceptance rate under null hypothesis. In other words the α value here is 0.0435, which is not sufficiently close to 0.0185 since the difference is 0.0250. To get a closer value, the critical value of 8.0 was adjusted to 10.50 which gave α_1=0.0126. This adjustment causes the other two acceptance rates for N(0, 9) and N(0, 25), to be larger.

With this adjustment all the tests, Chi-square, K-S and K-$S_{(g)}$ are at (nearly) the same α level, so comparison of their powers are easily performed.

For the U(10, 17) case the power of the Chi-square and the K-$S_{(g)}$ are 0.2208 and 0.2426 and for U(10, 13) both are equal to 1.00. It means that in the U(10, 17) case the K-$S_{(g)}$ test appears to be more powerful than the Chi-square test, whereas in the U(10, 13) case both tests have the same power. For the Exponential case with parameter 6, 12 and 15 the power of the K-$S_{(g)}$ test turns out to be more powerful than the Chi-square test in these Exponential and Normal cases. Looking at the results of another simulation, using three samples, generated from U(0, 18), U(0, 20) and U(0, 22) distributions, where the null hypothesis to be tested was the Exponential distribution with mean λ=9, which are also tabulated in table (5.1), the K-$S_{(g)}$ appears to be more powerful than the Chi-square test.

For the Normal case, the Chi-square test turns out to be more powerful than the K-$S_{(g)}$, even though the expected frequencies in the first and fourth groups were very low, only 2.220 and 2.218. Of course, the Chi-square distribution in these applications is only an approximation. This approximation gives better result for larger samples. It is a rule of thumb that the approximation can be used with confidence as long as every expected frequency is at least equal to 5. This rule should not be considered inflexible however. It appears to be a conservative value and the Chi-square approximation was reasonably accurate in this study, even for expected cell frequencies as small as 1.5 (6).

Based on the results of the simulation for Exponential cases where two of the four groups have low expected frequencies (namely 3.0315 and 2.2350), the approximation appears reasonably accurate. The estimated value of α_1 obtained from this simulation is α_1=0.0316 which is close to the "target," α_1=0.0328. By contrast, for the Normal case, the approximation is not so accurate. Thus, it was adjusted to get a closer value, as mentioned before. This inaccurate approximation may be caused by the fact of small expected frequencies in the tail.

Table (5.2) shows results of yet another simulation. By keeping the variance constant and letting the means vary, again the K-$S_{(g)}$ test appears to be more powerful than the Chi-square test.

In all cases, without grouping the data, in other words, in the continuous case, the ordinary K-S test is the most powerful test. Existing tables for the K-S distribution indicate the critical value δ=.3 corresponds to an α level somewhere between .10 and .15. To get the same, or at least close to the critical level used for the K-$S_{(g)}$, for example 0.0328 for the Exponential case, d=.3 should be increased. In other words, the value of δ for continuous case would be greater. Again, grouping the data and holding the same α level requires the critical value of the K-S test to be smaller than that tabulated.

Random sample	d	Number of group	$P[D_n \geq d] = \alpha$	Critical value of Chi-square associated with α	Null Hypothesis	Acceptance of K-S		
						χ^2	group	continuous
N(0,1)	.3	4	.0185	10.50*	N(0,1)	.9874	.9823	.8721
N(1,1)	.3	4				.2847	.2185	.1640
N(2,1)	.3	4				0	0	0
N(3,1)	.3					0	0	0

* has been adjusted

Table 5.2

VI. SUMMARY

1. Based on the results of the simulations, grouping the data causes the K-S statistic to be stochastically smaller than that for the ordinary tabulated K-S statistic. For continuous underlying distribution functions the K-S test is the most powerful of the three test considered.
2. It is hard to draw a general conclusion related to the relative powers of the $K\text{-}S_{(g)}$ and Chi-square tests. In some cases the $K\text{-}S_{(g)}$ test is more powerful than the Chi-square test; in other cases the reverse is true. It is suggested that the relative powers be investigated in more detail. Further investigation is also needed to determine rules of thumb for the appropriate number of groups to be used.
3. For small samples, the ordinary table of the K-S statistic cannot be used for grouped data. The extension of the K-S test suggested herein, should be used when the data have been grouped. Unfortunately, for sample sizes larger than 30 the calculation becomes tedious.
4. It would be very worthwhile to program a subroutine of Conover's procedure. The availability of the subroutine would enable this test to be used without time consuming calculations, and also possibly make the test available for larger sample sizes.

5. It is suggested that the possibility of developing a "quick and dirty" modification of the K-S table for use in discrete and grouped cases, be investigated.
6. This simulation also demonstrated the adequacy of χ^2 approximation, even when expected cell frequencies fell substantially below 5.
7. Care should be taken in computing D_n, the maximum vertical distance between $F_o(x)$ and $S_n^g(x)$. This maximum distance is equal to the greatest value of $\left|F_o(b_i) - S_n^g(b_i)\right|$ where b_i ranges over the set of class boundaries.

BIBLIOGRAPHY

1. Breiman, Leo; "Statistics, With a View Toward Application," Houghton Mifflin Company 1973.

2. Conover, W.J.; "Kolmogorov Goodness-of-Fit Test for Discontinuous Distribution," Journal of the American Statistical Association, Vol. 67, 1972, page 591-596.

3. D'Agostino, R.B. & Noether, G.E.; "On the Evaluation of the Kolmogorov-Statistics," The American Statistician, Vol. 27, 1973, page 81-82.

4. Davidson, T.G.; "A Modified Kolmogorov-Smirnov Test Applicable to Censored Sample," A thesis for the degree of Master of Science in Operation Research from the Naval Postgraduate School, 1971.

5. Feller W.; "On the Kolmogorov-Smirnov Times Theorem for Emperical Distribution," Annal of Mathematical Statistics, Vol. 21, 1950, page 177-189.

6. Gibbons, J.D.; "Nonparametric Statistical Inference," McGraw-Hill Book Company, 1971.

7. Massey, Jr. F.J.; "The Kolmogorov-Smirnov Test for Goodness-of-Fit," Journal of the American Statistical Association, Vol. 46, 1951, page 68-78.

8. Massey, Jr. F.J.; "A note on the Estimation of a Distribution Function by Confidence Limits," Annal of Mathematical Statistic, Vol. 21, 1950, page 116-119.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Department Chairman, Code 55 Department of Operation Research and Administrative Sciences Monterey, California 93940	2
4.	Assoc. Professor Donald R. Barr, Code 55Bn Department of Operation Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	2
5.	Chief of Staff of the Indonesian Army Headquarters of the Indonesian Army Jln. Veteran, Jakarta, Indonesia	1
6.	Biro Pusat Statistik Jln. Pos, Jakarta, Indonesia	1
7.	Akademi Ilmu Statistik Jln. Oto Iskandardinata, Jakarta-Timur Indonesia	1
8.	Professor Peter W. Zehna, Code 55Ze Department of Operation Research and Administrative Sciences Monterey, California 93940	1
9.	Letcol D. Suharto H.43 Nusaindah, Cijantung II Jakarta Timur, Indonesia	1

Thesis
D16165 Darmosiswoyo
c.1 Kolmogorov-Smirnov
test for grouped data.
160842